ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΑΞΙΝΟΜΗΣΗ ΣΥΝΑΙΣΘΗΜΑΤΩΝ ΜΕ ΒΑΣΗ ΤΗΝ ΕΠΟΠΤΕΥΟΜΕΝΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Διπλωματική Εργασία

Αλέξανδρος Κόντος

Επιβλέπων:

Σταμούλης Γεώργιος

Βόλος 2019

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

SENTIMENT CLASSIFICATION BASED ON SUPERVISED MACHINE LEARNING

Diploma Thesis

Alexandros Kontos

Supervisor:

George Stamoulis

Volos 2019

ΠΕΡΙΛΗΨΗ

Σκοπός αυτής της πτυχιακής εργασίας είναι η αναγνώριση συναισθήματος προτάσεων που εισάγονται από τον χρήστη μέσω της εποπτευόμενης μηχανικής μάθησης, δηλαδή θα προβλέπεται εάν οι προτάσεις αυτές είναι θετικές ή αρνητικές. Πιο συγκεκριμένα θα χρησιμοποιήσουμε μια βάση δεδομένων με σχόλια και αναρτήσεις από χρήστες του twitter την οποία και θα επεξεργαστούμε με διάφορες διαδεδομένες τεχνικές επεξεργασίας και μορφοποίησης δεδομένων κειμένου ώστε να έχουμε μια πιο ακριβής ανάλυση και πρόβλεψη. Με βάση αυτήν την επεξεργασία θα χρησιμοποιήσουμε πέντε διαφορετικούς classifier του οποίους θα αναλύσουμε σε θεωρητικό υπόβαθρο και θα συγκρίνουμε τις επιδόσεις τους στις προβλέψεις των δεδομένων. Τέλος μέσω μιας αυτοματοποιημένης διαδικασίας θα επιλέγεται ο classifier με την υψηλότερη απόδοση και με βάση αυτόν, θα γίνεται η πρόβλεψη συναισθήματος της πρότασης που εισάγεται από τον χρήστη.

ABSTRACT

The purpose of this thesis is the sentiment analysis of phrases given as input by a user, using supervised machine learning algoritmhs. More specificly, it will be stated if a phrase expresses possitive or negative sentiments. Furthermore, the data for training is provided by a database containing tweets, which are being proccessed with various and well known text formatting technincs in order to produce more accurate analysis and eventually sentiment foresight. Five different classifiers are implemented and tested for performance and accuracy. Finally, through an automated proccess the algorith with the best combination of performance and accuracy is selected in order to respond relatively to the sentiment of input phrases given by a user.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Η ραγδαία εξέλιξη της τεχνολογίας επέφερε και την εξέλιξη των μεθόδων της μηχανικής μάθησης. Οι υπολογιστές πλέον μπορούν να μάθουν από τα δεδομένα με λόγω της επαναλληπτικής πλευράς της επιστήμης αυτής. Καθώς εκτίθονται σε νέα δεδομένα οι υπολογιστές είναι σε θέση να προσαρμόζονται ανάλογα, μαθαίνοντας από προηγούμενος υπολογισμούς και έτσι μπορούν να παράγουν αξιόπιστα αποτελέσματα και να πάρουν αποφάσεις. Η πρόσφατη εξέλιξη στη τεχνολογία αυτή έχει δώσει την δυνατότητα εφαρμογής σύνθετων μαθηματικών υπολογισμών σε μεγάλες βάσεις δεδομένων με επαναλληπτικό τρόπο ταχύτερα [1].

Σε αυτή την διπλωματική εργασία θα επικεντρωθούμε στην εποπτευόμενη μηχανική μάθηση και την ανάλυση συναισθήματος. Στο δεύτερο κεφάλαιο παρουσιάζεται η ραγδαία αύξηση των μέσων κοινωνικών δικτύν και θα γίνει επεξήγηση των εννοιών της εξόρυξης δεδομένων και της ανάλυσης συναισθήματος. Στο κεφάλαιο 3 απεικονίζεται η αρχική κατάσταση του dataset που θα χρησιμοποιήσουμε και περιέχει τα βήματα επεξεργασίας των δεδομένων, παρουσιάζοντας επίσεις και το αντύκτιπο που έχει η εφαρμογή αυτών στο dataset μας. Ακόμα επεξηγούνται οι τεχνικές και οι αλγόριθμοι που θα χρησιμοποιούμε. Το κεφάλαιο 4 περιέχει το θεωρητικό υπόβαθρο των πέντε classifier που θα χρησιμοποιήσουμε και παρουσιάζονται οι αποδόσεις αυτών όταν εφαρμοστούν στο dataset που έχουμε ήδη επεξεργαστεί. Τέλος στο κεφάλαιο 5 περιγράφεται η διαδικασία αυτοματοποιημένης επιλογής του αποδοτικότερου classifier και αποτυπώνονται μερικά παραδείγματα πρόβλεψης συναισθήματος από προτάσεις που εισάγει ο χρήστης.

ΚΕΦΑΛΑΙΟ 2

ΜΕΣΑ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ ΚΑΙ ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

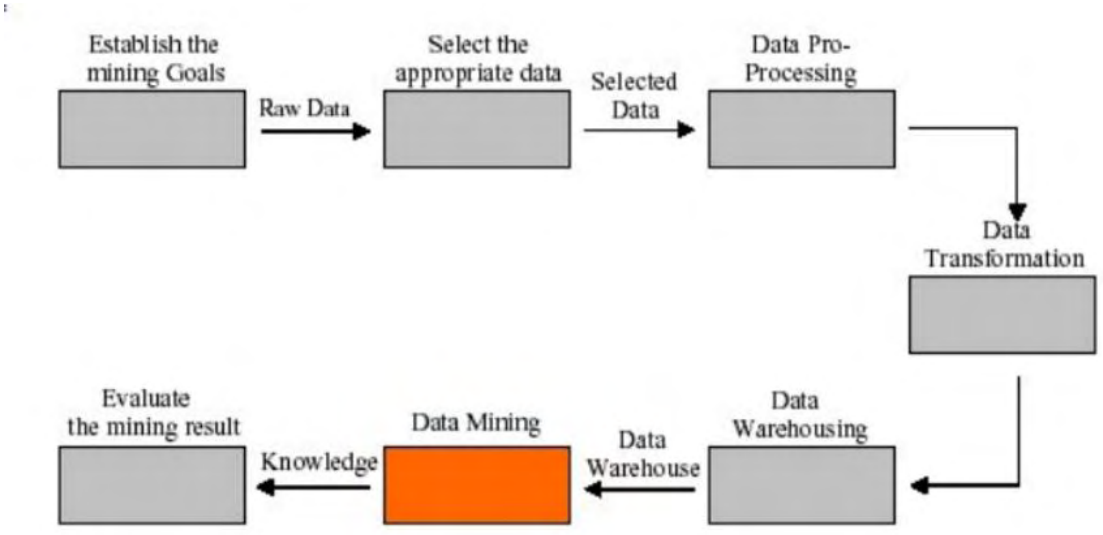
2.1 Μέσα κοινωνικής δικτύωσης και twitter

Από τις αρχές της πρηγούμενης δεκαετίας η ραγδαία εξέλιξη της τεχνολογίας και ειδικότερα η εξέλιξη των έξυπνων κινητών συνέβαλε στην εύκολη πρόσβαση στο διαδίκτυο από ανθρώπους όλων των ηλικιών. Προέκταση του γεγονότος αυτού είναι η πρόσβαση στη πληροφορία από αμέτρητες πηγές και η αναδιαμόρφωση των πολιτικών των εταιριών που ολοένα στρέφουν τις δραστηριότητές τους προς το διαδίκτυο. Ως εκ τούτου, άρχισαν να κάνουν την εμφάνισή τους και τα μέσα κοινωνικής δικτύωσης που με τη πάροδο του χρόνου έως και σήμερα μετρούν δισεκατομύρια χρήστες και πιο συγκεκριμένα σχεδόν το ένα τρίτο του πλυθησμού της γης έιναι εγγεγραμμένοι σε κάποια από αυτού του είδους τις πλατφόρμες. Μερικά από αυτά είναι το facebook, το instagram, το twitter και άλλα.

Σε αυτή τη διπλωματική εργασία κύριο λόγο θα έχει το twitter, ένα μέσο στο οποίο οι χρήστες έχουν τη δυνατότητα ενημέρωσης από κάθε είδους πηγή και κυρίως ο καθένας είναι ελεύθερος να αλληλεπιδρά, να διατυπώνει και να ανταλλάσει τις απόψεις του ακόμα και τις εμπειρίες του με άλλους χρήστες. Ένας ακόμη λόγος που είναι ευρέως διαδεδομένο το twitter, είναι το γεγονός πως ένα μεγάλο ποσοστό των ανθρώπων που είναι ιδίετερα γνωστοί στο κοίνο για οποιονδήποτε λόγο, όπως για παράδειγμα αρκέτοι πολιτικοί εκφράζονται μέσω αυτού και απαριθμούν εκατομμύρια ακολόθους. Επίσεις στον τομέα του marketing, τα social media πλέον αποτέλουν αναπόσπαστο εργαλείο, είτε ανάλυσης και λήψης feedback, είτε διαφήμισης.

2.2 Εξόρυξη δεδομένων και ανάλυση συναισθήματος

Η εξόρυξη δεδομένων είναι η διαδικασία ταξινόμησης μέσω μεγάλων βάσεων δεδομένων για τον προσδιορισμό μοτίβων και τη δημιουργία σχέσεων για την επίλυση προβλημάτων μέσω της ανάλυσης δεδομένων και τη πρόβλεψη μελλοντικών τάσεων.. Αρχικά πρέπει να καθοριστούν οι στόχοι της εξόρυξης, να επιλεγούν τα κατάλληλα δεδομένα, να γίνει η κατάλληλη επεξεργασία και μεταμόρφωση και τέλος να αξιολογηθούν τα αποτελέσματα όπως φαίνεται στο παρακάτω σχήμα [1].



Εικόνα 2.1 : Στάδια της διαδικασίας της εξόρυξης δεδομένων [2]

Ώσον αφορά την επιστήμη των δεδομένων και την μηχανική μάθηση το twitter κατατάσσεται σε μια από τις κορυφέες πλατφόρμες που παρέχουν μεγάλη ποσότητα δεδομένων. Η εξόρυξη δεδομένων μέσω του twitter ουσιαστικά αφορά την συλλογή μη επεξεργασμένων πληροφοριών, όπως είναι οι αναρτήσεις και ο σχολιασμός tweets τα οποία δημοσιεύονται σ' αυτό, οι οποίες μπορούν να χρησιμοποίηθούν με πολλούς τρόπους όπως η δημιουργία κοινωνικών προφίλ και ο προσδιορισμός μελλοντικών τάσεων.

Η ανάλυση συναισθήματος είναι η διαδικασία όπου καθορίζεται ένα ένα κείμενο γραφής παρουσιάζει θετικό, αρνητικό ή ουδέτερο συναίσθημα, χρησιμοποιούνται τεχνικές φυσικής επεξεργασίας γλώσσας και μηχανικής μάθησης και αποδίδονται σταθμισμένες βαθμολογίες στις εκάστοτε προτάσεις ή φράσεις. Αυτή η διαδικασία χρησιμοποιείται ευρέως από μεγάλες επιχειρήσεις με σκοπό την κατανόηση και την εκτιίμηση της κοινής γνώμης παρακολουθώντας και κατανοώντας έτσι τις εμπειρίες των πελατών τους στα προιόντα τους.

Οι τεχνικές που ακολουθούνται για την ανάλυση συναισθήματος είναι κυρίως ο διαχωρισμός κάθε εγγράφου στα συστατικά του μέρη, δηλαδή σε προτάσεις, φράσεις και λέξεις. Κατόπιν προσδιορίζεται το συναίσθημα που φέρει κάθε φράση και λέξη και ορίζεται μια βαθμολογία συναισθήματος με μια κλίμακα τριων τιμών για τον διαχωρισμό των τριών συναισθημάτων [3]

ΚΕΦΑΛΑΙΟ 3

DATASET ΚΑΙ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

3.1 Ανάλυση των δεδομένων

Η βάση δεδομένων που χρησιμοποιείται σ' αυτήν την εργασία προέρχεται από το σάιτ kaggle με τον τίτλο sentiment 140 και περιέχει σχόλια και αναρτήσεις από χρήστες του twitter. Η αρχική μορφή του παρουσιάζεται στην παρακάτω εικόνα.

	target	id	date	flag	user	text
0	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
1	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
2	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
3	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....
4	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew

Εικόνα 3.1 : Αρχική κατάσταση του dataset

Το dataset αποτελείται από 1.6 εκατομμύρια γραμμές και έξι στήλες, με την στήλη target να αποτελεί τη στήλη με το συναίσθημα και έχει τιμές, 0 για αρνητικό συναίσθημα και 4 για θετικό συναίσθημα. Στη περιγραφή του dataset στο σάιτ της kaggle αναφέρεται ότι υπάρχει και η τιμή 2 για ουδέτερο συναίσθημα, ωστόσο δεν παρατηρείται πουθενά όπως φαίνεται

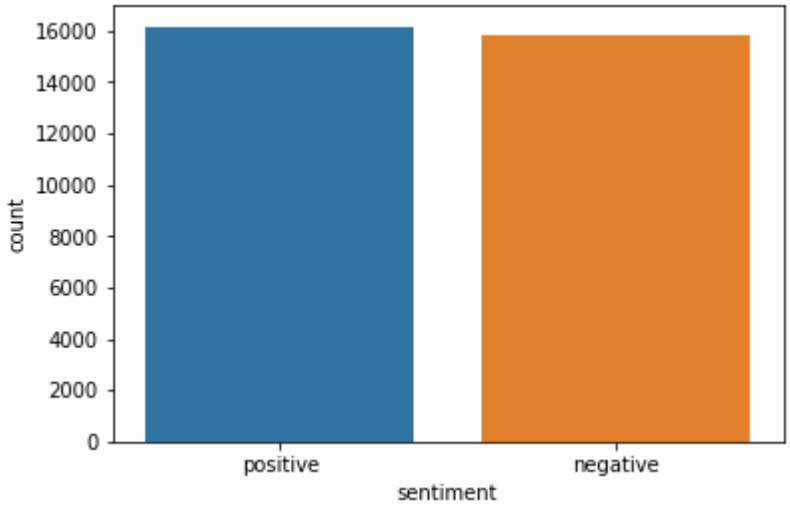
στην εικόνα 3.3. Επίσεις έχουμε τη στήλη id με έναν αριθμό για κάθε γραμμή του dataset, τη στήλη date με την ημερομηνία του tweet και τη στήλη user με το όνομα του χρήστη για κάθε tweet. Η στήλη flag περιγράφει κατά πόσο υπάρχει ειρωνία σε κάθε σχόλιο παρόλα αυτά υπάρχει μόνο η τιμή NO_QUERY για όλο το dataset. Τέλος έχουμε τη στήλη text που περιέχει τα tweets.

	sentiment	text
1281582	positive	@dilipm here's the deal - can isos make up for...
669476	negative	exam week! :| hectic!?
238918	negative	my day is only half over. sad.
942179	positive	at the bus stop chilli with tyler jones
1586834	positive	wahey - threads: 990, posts: 3,030, members: 2...

Εικόνα 3.2 : Απλοποιημένη μορφή του dataset

Από τις έξι αυτές στήλες είναι χρήσιμες μόνο οι δυο για την ανάλυση, οι στήλες target και text τις οποίες και απομονώνουμε. Την στήλη target την μετονομάζουμε σε sentiment και αλλάζουμε τις τιμές της από 0 σε negative και από 4 σε positive. Ακόμη θα ήταν αδύνατο να δουλέψουμε με 1.6 εκατομμύρια γραμμές καθώς δεν υπάρχει ο απαραίτητος εξοπλισμός και έτσι κρατάμε ένα τυχαίο δείγμα 32000 γραμμών από το σύνολο του dataset, αριθμός που μας επιτρέπει να δουλέψουμε χωρίς μεγάλες αναμονές για την επεξεργασία του dataset. Η τελική μορφή του φαίνεται στην εικόνα 3.2 και το

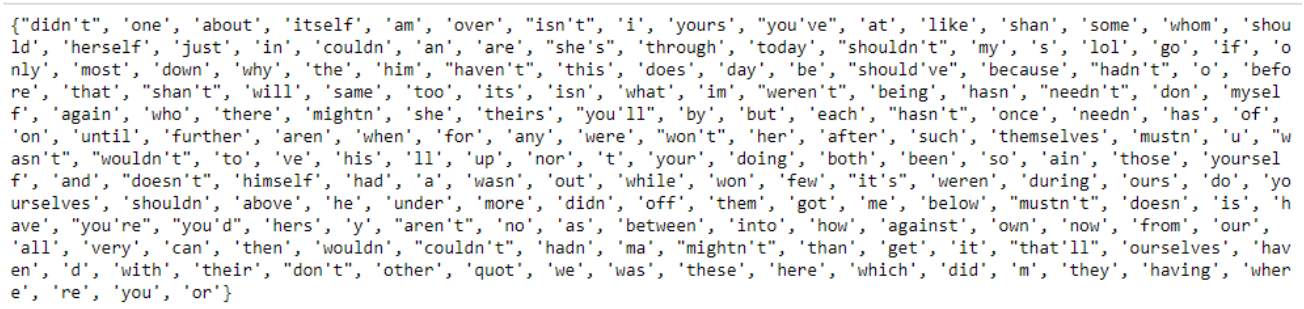
σύνολο των τιμών σην εικόνα 3.3 όπου έχουμε λίγο πάνω από 16000 θετικά σχόλια και λίγο κάτω από 16000 αρνητικά σχόλια.
.



Εικόνα 3.3 :σύνολο τιμών της στήλης target

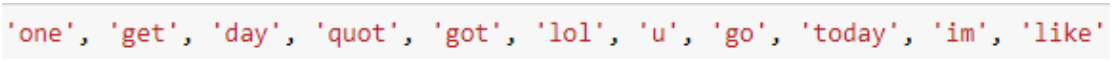
3.2 Προεπεξεργασία των δεδομένων

Αφού κρατήσουμε μόνο τις στήλες που χρειαζόμαστε, θα δημιουργήσουμε μια κλάση με όνομα preprocessing στην οποία στέλνουμε το dataset όπου και θα γίνει ολόκληρη η διαδικασία της προεπεξεργασίας.Αρχικά ‘καθαρίζουμε’ τα δεδομένα από τιμές που δεν υπάρχουν ή είναι μη έγκυρες και τις αντικαθιστούμε με κενά και μετατρέπουμε τυχόν κεφαλαίους χαρακτήρες σε πεζούς**.** Επίσεις αφαιρούμε όλα τα url και τα tags που περιέχουν τον χαρακτήρα @ ακολουθούμενο από όνομα χρήστη όπως φαίνεται στη πρώτη γραμμή της εικόνας 3.2. Στη συνέχεια κάνουμε εγκατάσταση ένα πακέτο της python, το nltk το οποίο θα μας βοηθήσει για τη συνέχεια της διαδικασίας. Πιο συγκεκριμένα μέσω αυτού του πακέτου θα καλέσουμε τη συνάρτηση word_tokenize η οποία σπάει τις προτάσεις σε λέξεις. Αυτή η διαδικασία θα μας χρησιμεύσει σε επόμενο στάδιο της επεξεργασίας όπου θα γίνει μετατροπή των λέξεων σε αριθμούς για την αλγοριθμική ανάλυση. Επόμενο βήμα είναι να χρησιμοποιήσουμε ακόμα μια συνάρτηση από το nltk την stopwords. Stopwords είναι λέξεις οι οποίες δε προσφέρουν κάποια χρησιμότητα στην αλγοριθμική ανάλυση που θα γίνει παρακάτω, αντιθέτως ‘φορτώνουν’ το σύστημα με παραπάνω άσκοπη δουλειά.



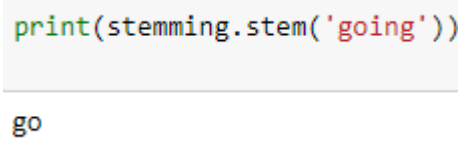
Εικόνα 3.4 : Προεπιλεγμένα stopwords

Η συγκεκριμένη συνάρτηση έχει από προεπιλογή κάποιες λέξεις(εικόνα 3.4) και προσθέτουμε σ' αυτές μερικές ακόμα που πιστεύουμε πως δεν θα έχουν κάποια χρησιμότητα(εικόνα 3.5)



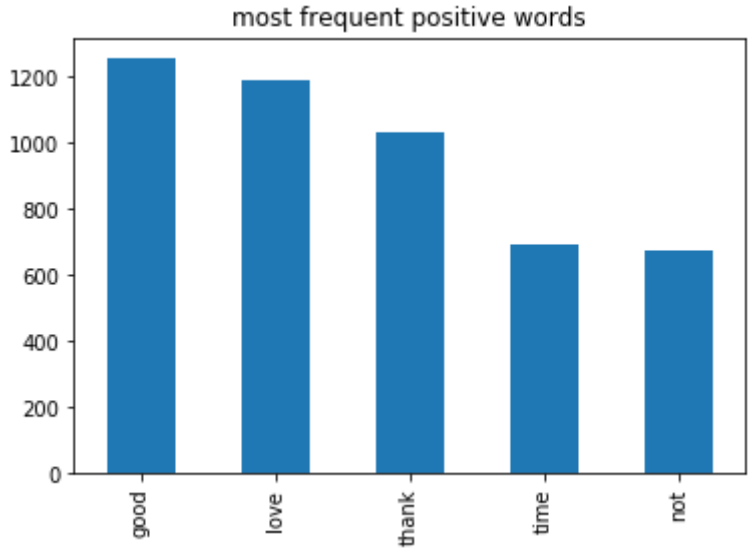
Εικόνα 3.5 : Προστιθέμενα stopwords

Κατόπιν θα χρησιμοποιήσουμε τη τεχνική stemming για την ομαδοποίηση ίδιων λέξεων με διαφορετικό επίθεμα, ώστε να λογίζονται ως ίδιες από το σύστημά μας. Για παράδειγμα η λέξη going μετατρέπεται σε go(εικόνα 3.6)

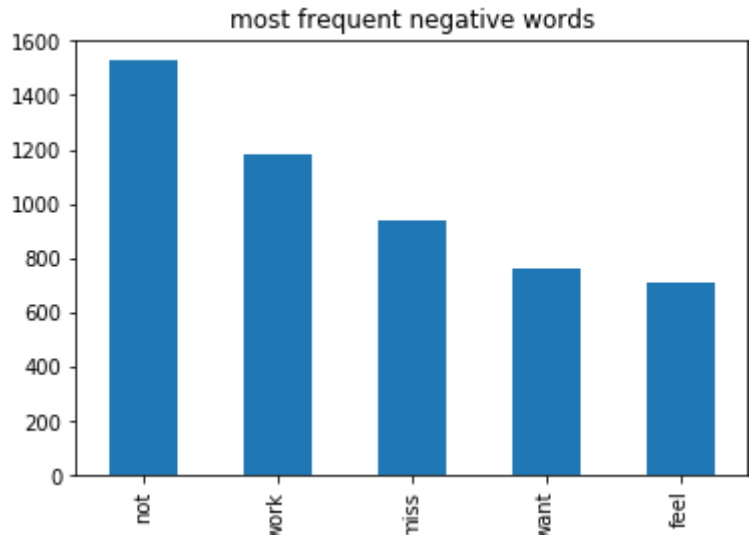


Εικόνα 3.6 : Παράδειγμα συνάρτησης stemming

Στην εικόνα 3.9 απεικονίζεται το τελικό αποτέλεσμα της προεπεξεργασίας, ενώ στις εικόνες 3.7 και 3.8 φαίνονται οι περισσότερο εμφανιζόμενες θετικές και αρνητικές αντίστοιχα λέξεις. Παρατηρούμε ότι υπάρχουν κοινές λέξεις και στα δυο πεδία, γεγονός που θα επηρεάσει αρνητικά την απόδοση των classifier όπως θα δούμε παρακάτω.



Εικόνα 3.7 : Πιο συχνά εμφανιζόμενες θετικές λέξεις



Εικόνα 3.8 : Πιο συχνά εμφανιζόμενες αρνητικές λέξεις

	sentiment	text
1281582	positive	deal iso make fill flash
669476	negative	exam week hectic
238918	negative	half sad
942179	positive	bu stop chilli tyler jone
1586834	positive	wahey thread post member lot mileston forum we...

Εικόνα 3.9 : Αποτέλεσμα προεπεξεργασίας

Επίσεις θα πρέπει να μετατρέψουμε τις προεπεξεργασμένες λέξεις σε αριθμούς ώστε να κατανοούνται από τον υπολογιστή. Αυτό θα πραγματοποιηθέι με τη συνάρτηση από την sklearn, TfidfVectorizer. Αφού έχουμε μετατρέψει κάθε πρόταση σε ένα σύνολο από λέξεις, η συνάρτηση αυτή αποδίδει σε κάθε λέξη του συνόλου έναν αριθμό και όσο προχωράει αυτή η διαδικασία όσο πιο συχνά εμφανίζεται μια λέξη τόσο αυξάνεται και ο αποδιδόμενος αριθμός της. Στην παρακάτω εικόνα παρουσιάζεται το αποτέλεσμα της συγκεκριμένης διαδικασίας στο περιεχόμενο της εικόνας 3.9.

```
(0, 4388)     0.31563512808807603
(0, 2642)     0.5968682050058217
(0, 2605)     0.5358309809398014
(0, 1836)     0.5069595364869293
(1, 7892)     0.39522947687232945
(1, 3296)     0.7786919691000945
(1, 2440)     0.48727043607245696
(2, 6149)     0.5956358529583825
(2, 3175)     0.8032545864609427
(3, 7544)     0.5260923828214638
(3, 6864)     0.3318268632059832
(3, 3865)     0.4887453636347273
(3, 1305)     0.4601549227741958
(3, 991)      0.403116799655316
(4, 7902)     0.3074324412064719
(4, 7224)     0.44006136917767713
(4, 5595)     0.2948246297510305
(4, 4625)     0.468249447709768
(4, 4545)     0.4017678878969923
(4, 4301)     0.2722205766681831
(4, 2729)     0.41246902884482595
(5, 8197)     0.2567529736659077
(5, 5945)     0.3546466933579547
(5, 5849)     0.2407216078208423
(5, 5798)     0.3203185812531789
```

Εικόνα 3.10 : Αποτέλεσμα συνάρτησης μετατροπής TFidfVectorizer

Τέλος παίρνουμε τις δυο στήλες του dataset και της φορτώνουμε σε δυο μεταβλητές, τη στήλη με το text στο X και τη στήλη με το συναίσθημα στο Y και το τα χωρίζουμε σε δυο κομμάτια, το train και το test με ποσοστό 0.8 και 0.2 αντίστοιχα. Με αυτή τη διαδικασία τα δεδομένα μας είναι έτοιμα να περαστούν από τους classifiers για να δοκιμάσουμε τα ποσοστά προβλέψεών τους.

ΚΕΦΑΛΑΙΟ 4

ΑΝΑΛΥΣΗ ΑΛΓΟΡΙΘΜΩΝ(CLASSIFIERS ΚΑΙ ΑΞΙΟΛΟΓΗΣΗ ΜΕΤΡΗΣΕΩΝ)

Σε αυτό το κεφάλαιο θα γίνει ανάλυση και σύγκριση των αποτελεσμάτων από πέντε διαφορετικούς classifier, των logistic regression, decision tree, naïve bayes, svm και random forest. Αρχικά θα επεξηγήσουμε τις παραμέτρους του confusion matrix και την τεχνική του cross-validation και για να γίνουν κατανοήτα τα αποτελέσματα της αλγοριθμικής ανάλυσης.

- **Accuracy :** Δείχνει το ποσοστό σωστών προβλέψεων που έκανε ο classifier

- **Precision :** Το ποσοστό των μεταβλητών που είναι θετικές και προβλέφθηκαν σωστά.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- **Recall :** Το ποσοστό σωστών προβλέψεων των μεταβλητών

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- **F1-score :** Συνδιάζει το precision και το recall για καλύτερη μέτρηση

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

- **Cross-validation :** Είναι μια τεχνική αξιολόγησης των αποτελεσμάτων της ανάλυσης και πως αυτά γενικεύονται σε ένα ανεξάρτητο dataset. Ουσιαστικά υπολογίζει κατά πόσο είναι ακριβές ένα μοντέλο πρόβλεψης

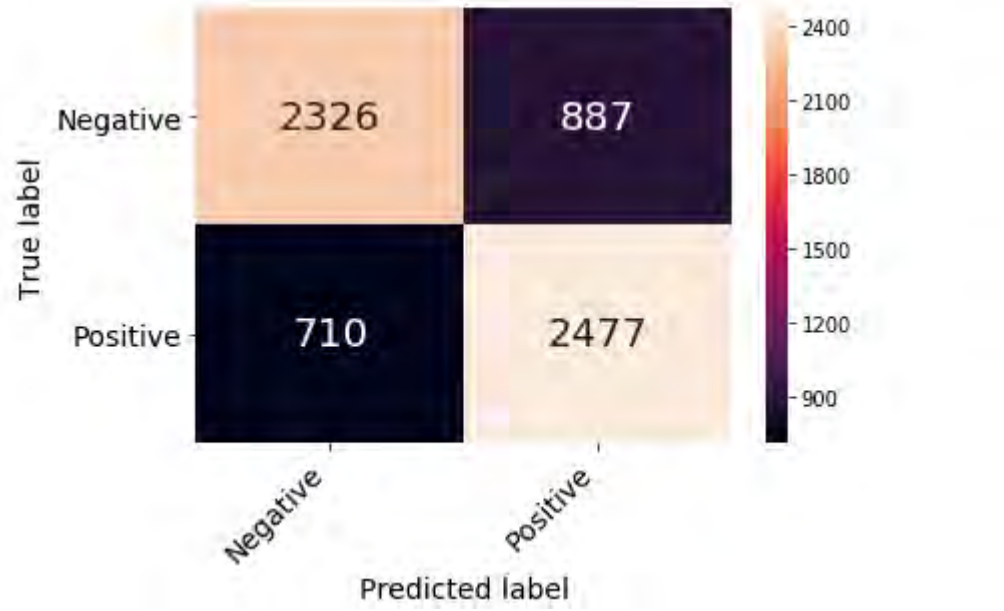
1. Logistic Regression

Είναι μια στατιστική μέθοδος που αναλύει ένα dataset όπου υπάρχουν μια ή περισσότερες ανεξάρτητες μεταβλητές που καθορίζουν το αποτέλεσμα το οποίο μετράται με μια διχοτόμο μεταβλητή στην οποία δυο πιθανά αποτελέσματα, είναi δηλαδή δυαδική και κωδικοποιείται ως 1 για αληθές η 0 ως ψευδές. Στόχος αυτής της μεθόδου είναι να βρεθεί το καλύτερο μοντέλο που περιγράφει τη σχέση ανάμεσα στην διχοτόμο μεταβλητή και το σύνολο από τις ανεξάρτητες μεταβλητές [4]. Το μαθηματικό μοντέλο :

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Όπου ℓ η λογαριθμική απόδοση, b η βάση του λογαρίθμου και bi οι παράμετροι του μοντέλου. Στην παρακάτω εικόνα βλέπουμε τα αποτελέσματα του logistic regression στο dataset μας.

```
Accuracy of logistic regression classifier on test set: 0.75
[0.75195251 0.75132771 0.75945017 0.7553125  0.7484375  0.739375
 0.74875    0.75179744 0.75304783 0.76492654]
              precision    recall  f1-score   support

    negative       0.77      0.72      0.74      3213
    positive       0.74      0.78      0.76      3187

avg / total        0.75      0.75      0.75      6400
```



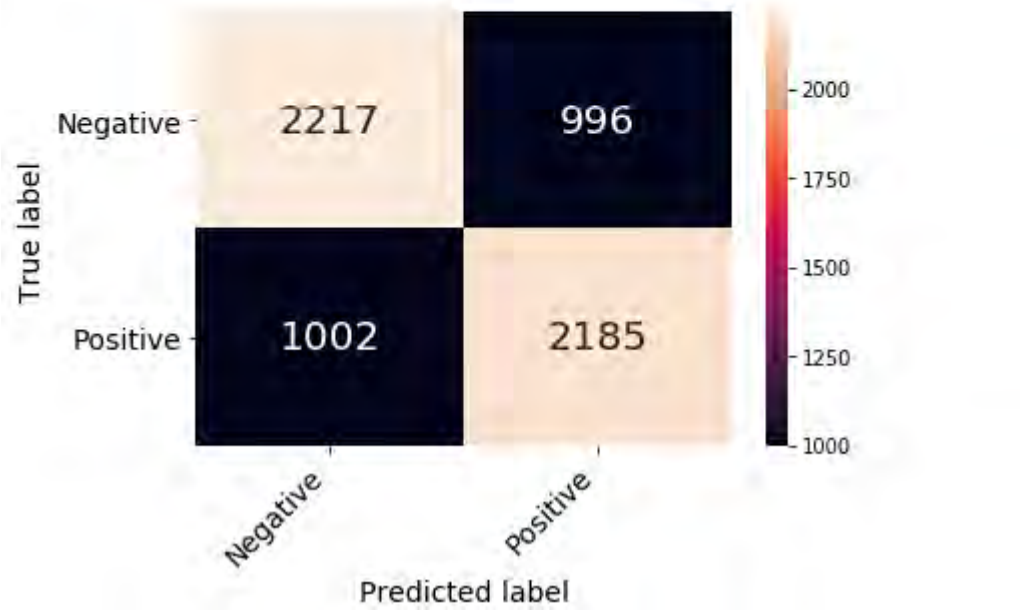
Εικόνα 4.1 : Logistic Regression

Παρατηρούμε πως η ακρίβεια αυτού του classifier είναι 0.75, γεγονός που επιβεβαιώνεται και από το cross-validation(κόκκινο πλαίσιο), υπολογισμένο με cv=10 δηλαδή το έχουμε υπολογίσει σε δέκα διαφορετικές πτυχές. Έχουμε 2326 αρνητικές και 2477 θετικές προβλέψεις που είναι σωστές, 710 λανθασμένες αρνητικές προβλέψεις που ήταν θετικές και 887 λανθασμένες θετικές προβλέψεις που ήταν αρνητικές. Συνολικά έχουμε 4.803 σωστές και 1.597 λάθος προβλέψεις σε 6.400 μεταβλητές.

2. **Decision Tree**

Είναι ένας σχετικά απλός αλγόριθμος με δομή δέντρου όπου κάθε κόμβος αντιπροσωπεύει ένα χαρακτηριστικό, ο κλάδος έναν κανόνα απόφασης και κάθε φύλλο το αποτέλεσμα. Ο υψηλότερος κόμβος αποτελέι τον κόμβο ρίζας και διαχωρίζεται με βάση την τιμή του χαρακτηριστικού, αυτό συνβαίνει επαναλαμβανόμενα σχηματίζοντας έτσι το δέντρο που βοηθά στην λήψη αποφάσεων [5].

```
Accuracy of decision tree classifier: 0.69
cv_scores mean:
[0.68759763 0.69946892 0.70071853 0.6934375  0.680625   0.6803125
 0.6934375  0.69052829 0.69459206 0.69709284]
             precision    recall  f1-score   support

   negative       0.69      0.69      0.69      3213
   positive       0.69      0.69      0.69      3187

avg / total       0.69      0.69      0.69      6400
```



Εικόνα 4.2 : Decision Tree classifier

Φαίνεται πως ο αλγόριθμος αυτός έχει λιγότερη απόδοση στο dataset μας με ποσοστό 0.70 που επιβεβαιώνεται από το cross-validation με συνολικά 4.402 σωστές προβλέψεις και 1998 λανθασμένες στο σύνολο των 6.400 μεταβλητών.

3. **Multinomial Naïve Bayes**

Χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης κειμένων, βασίζεται στη συχνότητα εμφανίσεως των λέξεων μέσα στο κείμενο και είναι ιδανικό για προβλήματα ταξινόμησης συναισθημάτων. Το θεώρημα του Bayes [6] :

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Όπου y η πιθανότητα να συμβεί ένα γεγονός και X = (X1,X2,.....Xn) οι παράμετροι που το επηρεάζουν.

```
Accuracy of Naive bayes classifier on test set: 0.74
cv_scores mean:
[0.73976882 0.74039363 0.73570759 0.7396875  0.7275     0.73375
 0.7384375  0.74116912 0.73929353 0.74773367]
             precision    recall  f1-score   support

   negative       0.74      0.74      0.74      3213
   positive       0.74      0.74      0.74      3187

avg / total       0.74      0.74      0.74      6400
```

Εικόνα 4.3 : Naïve Bayes

Σ’ αυτή την περίπτωση έχουμε ποσοστό προβλέψεων 0.74 που επίσεις επιβεβαιώνεται από το cross-validation με συνολικά 4.727 σωστές και 1.673 λάθος προβλέψεις σε 6.400 μεταβλητές.

4. Linear SVM (Support Vector Machine)

Ο αλγόριθμος SVM βρίσκει ένα διαχωριστικό υπερπλαίσιο με το μέγιστο περιθώριο μεταξύ δυο κλάσεων δεδομένων [7]. Έχοντας ένα σετ από ζεύγη με (xi, yi), xi ∈ Rn και yi ∈ {-1, 1}, i=1,….,l ο svm λύνει το παρακάτω χωρίς περιορισμούς πρόβλημα βετιστοποίησης

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\xi(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i),$$

Με ξ(w, b; xi , yi) η λανθάνουσα συνάρτηση και C>=0 παράμετρος ποινής. Δυο πολύ κοινές συναρτήσεις είναι οι ακόλουθες:

$$\max(1 - y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b), 0) \text{ and } \max(1 - y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b), 0)^2,$$

Φ είναι η συνάρτηση χαρτογράφησης των δεδομένων σε μεγαλύτερο χώρο διαστάσεων. Η συνάρτηση απόφασης για x είναι

$$f(x) = sgn(w^T \phi(x) + b)$$

Ενώ η συνάρτηση πυρήνα για τον linear svm δίνεται από τον τύπο:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j.$$

Η εφαρμογή του linear svm στο dataset μας:

```
Accuracy of SVM classifier: 0.74
cv_scores mean:
[0.74789128 0.74570447 0.7503905  0.7484375  0.739375   0.73875
 0.7496875  0.75304783 0.74210691 0.75554861]
             precision    recall  f1-score   support

   negative       0.76      0.72      0.74      3213
   positive       0.73      0.77      0.75      3187

avg / total       0.74      0.74      0.74      6400
```

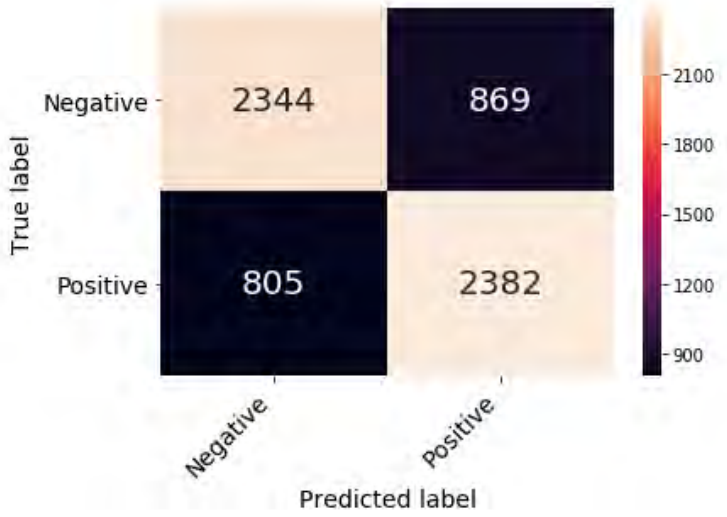
Εικόνα 4.4 : Linear SVM

Σε αυτή την περίπτωση παρατηρούμε πως έχουμε το ίδιο σκορ με τον Naïve Bayes classifier και ταυτόχρονα βρίσκεται πολύ κοντά με αυτό του Linear Regression. Έχουμε συνολικά 4.762 σωστές προβλέψεις και 1.638 λάθος.

5. Random Forest classifier

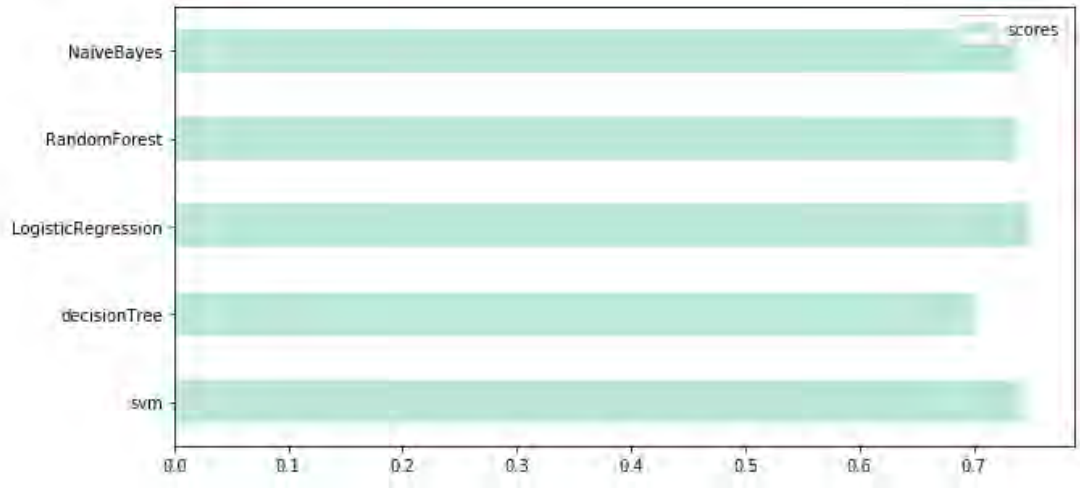
Ο αλγόριθμος Random Forest είναι μια εξελιγμένη μορφή του Decision Tree, ουσιαστικά αποτελείται από πολλαπλά decision trees. Έχοντας ένα training set $X=\chi1,\chi2,...,\chi n$ με απόκριση Y=y1,y2,..,yn και εφαρμόζοντας την μέθοδο baggind επανειλημμένα, επιλέγεται ένα τυχαίο δείγμα από το training set και εφαρμόζει δέντρα σε αυτό το δείγμα.

```
Accuracy of Random forest classifier: 0.74
cv_scores mean:
[0.74226804 0.7494533  0.74664167 0.7309375  0.7353125  0.7440625
 0.7415625  0.74617068 0.73460456 0.75804939]
              precision    recall  f1-score   support

    negative       0.74      0.73      0.74      3213
    positive       0.73      0.75      0.74      3187

 avg / total       0.74      0.74      0.74      6400
```



Εικόνα 4.5 : Random Forest

Έχουμε 0.74 ποσοστό σωστών προβλέψεων με 4.726 σωστές και 1.674 λάθος προβλέψεις, επίσεις ένα από τα υψηλότερα σκορ που παρατηρήσαμε απ αυτούς του αλγόριθμους. Στην εικόνα 4.6 παρουσιάζεται ένα συνολικό διάγραμμα με τα ποσοστά όλων των classifier.



Εικόνα 4.6 : Διάγραμμα classifiers

ΚΕΦΑΛΑΙΟ 5

ΥΛΟΠΟΙΗΣΗ

Στο προηγούμενο κεφάλαιο χρησιμοποιήσαμε πέντε διαφορετικούς classifier και συγκρίναμε τα αποτελέσματα του. Σε αυτό το κεφάλαιο παρουσιάζεται η αυτοματοποιημένη διαδικασία σύγκρισης των classifier και αποθήκευσης αυτού με το υψηλότερο ποσοστό πρόβλεψης σε μια νέα μεταβλητή ώστε να χρησιμοποιηθέι στην πρόβλεψη συναισθήματος της πρότασης που εισάγει ο χρήστης. Αρχικά να σημειωθέι πως ήδη γνωρίζουμε τον classifier με την υψηλότερη απόδοση και θα μπορούσαμε να τον επιλέξουμε κατευθείαν για τις προβλέψεις μας παρκάτω, ωστόσο σκοπός αυτού του κεφαλαίου ειναι η παρουσίασης της αυτοματοποιημένης διαδικασίας αυτής.

```
Best score is: 0.743  from classifier: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

Εικόνα 5.1 : Εκτύπωση αποδοτικότερου classifier

Με λίγα λόγια την υλοποίηση θα την αναθέσουμε στον υπολογιστή και στο πρόγραμμά μας. Θα χρειαστούμε την συνάρτηση GridSearchCV της python η οποία και θα εκτελέση αυτή τη διαδικασία, οπότε θα αναλύσουμε τις παραμέτρους της. Τα ορίσματα που θα εισάγουμε σε αυτή τη συνάρτηση είναι ένα pipeline και ακόμη μια μεταβλητή στην οποία εισάγουμε τους

estimators από τους classifiers μας. Όσων αφορά το pipeline μετά από δοκιμές παρατηρήθηκε πως είναι αναγκαίο και έτσι του εισάγουμε έναν από τους estimators των classifier χωρίς να έχει κάποια σημασία ποιοόν θα επιλέξουμε. Τέλος ορίζουμε μια νέα μεταβλητή στη οποία καλούμε την GridSearchCV με τα παραπάνω ορίσματα και έχουμε το επιθυμητό αποτέλεσμα όπως φαίνεται στην εικόνα 5.1 .

Στις παρακάτω εικόνες παρουσιάζονται μερικές δοκιμές προτάσεων και η πρόβλεψη συναισθήματός τους.

```
Enter your sentence: nice weather
Starting processing data...
           text
0  nice weather
  (0, 7933)     0.750673515605258
  (0, 4948)     0.6606733481599227
['positive']
```

Εικόνα 5.2 : Παράγειγμα 1

Εισάγοντας την πρόταση nice weather θα πρέπει να επαναλλάβουμε την διαδικασία που κάναμε για το dataset μας ώστε να γίνει η ίδια επεξεργασία στην πρόταση που εισάγουμε. Όπως φαίνεται στην εικόνα 5.1 αυτή η διαδικασία πραγματοποιείται, μέσω της κλάσσης που έχουμε δημιουργήσει στον κώδικα και στον οποίο πραγματοποιείται ολόκληρος ο ‘καθαρισμός’ των δεδομένων, η ίδια διαδικασία δηλαδή που επεξηγήθηκε στο κεφάλαιο 3.

```
Enter your sentence: bad weather
Starting processing data...
          text
0  bad weather
  (0, 7933)     0.763679635623156
  (0, 507)      0.6455953950691435
['negative']
```

Εικόνα 5.3 : Παράδειγμα 2

Εισάγωντας την πρόταση bad weather λαμβάνουμε επίσεις σωστή πρόβλεψη.

```
Enter your sentence: i wasn' t going to be prepared but it was all good
Starting processing data...
             text
0  go prepar good
  (0, 5652)     0.8095960871403125
  (0, 2995)     0.39615549533234395
  (0, 2977)     0.4331454711815428
['positive']
```

Εικόνα 5.4 : Παράδειγμα 3

Ένα ακόμη παράδειγμα είναι η πρόταση I wasn’ t going to be prepared but it was all good. Παρατηρούμε αρχίκα ότι μέσω της αφαίρεσης των stopwords έχουμε μόνο τις λέξεις going prepared και good, οι οποίες λόγω του stemming μετατρέπονται σε go, prepar και το good παραμένει ως έχει. Ακόμη παρατηρούμε και την επίδραση της συνάρτησης

TfidfVectorizer, η οποία μετατρέπει τις τρεις αυτές λέξεις σε αριθμούς για την κατανόησή τους από τον υπολογιστή και τέλος σωστά παίρνουμε τη πρόβλεψη για θετικό συναίσθημα από τον εκπαιδευμένο classifier μας με το υψηλότερο score.

Τέλος θα δοκιμάσουμε ένα παράδειγμα με λέξεις που βρίσκονται και στις πιο θετικά και στις πιο αρνητικά εμφανιζόμενες λέξεις όπως είναι οι not και love. Θα δοκιμάσουμε την πρόταση “I did not really love this song”.

```
Enter your sentence: i did not really love this song
Starting processing data...
                  text
0  not realli love song
  (0, 6740)     0.6327976476782253
  (0, 5909)     0.4938188086407976
  (0, 5025)     0.4036352588329123
  (0, 4315)     0.43907709932581973
['positive']
```

Εικόνα 5.5 : Παράδειγμα 4

Παρατηρούμε πως μας εμφανίζεται λανθασμένα θετικό συναίσθημα ενώ θα έπρεπε να τυπώνεται αρνητικό, αυτό ωφείλεται στο γεγονός πως ίδιες λέξεις εμφανίζονται και στις θετικές και στις αρνητικές λέξεις. Ειδικότερα για την λέξη not η οποία εμφανίζεται στις περισσότερο θετικές λέξεις αποδεικνύει τις δυσκολίες που υπάρχουν στο συγκεκριμένο dataset και εν μέρη δικιολογεί και τα accuracy score των classifier που χρησιμοποιήσαμε τα οποία δε ξεπερνούν το 75%.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] sas.com. [Ηλεκτρονικό]. Available: https://www.sas.com/en_us/insights/analytics/machine-learning.html.

[2] techopedia.com. [Ηλεκτρονικό]. Available: https://www.techopedia.com/definition/1181/data-mining.

[3] researchgate.net. [Ηλεκτρονικό]. Available: https://www.researchgate.net/figure/Stages-of-data-mining-process_fig1_27483416.

[4] lexalytics.com. [Ηλεκτρονικό]. Available: https://www.lexalytics.com/technology/sentiment-analysis.

[5] medcalc.org. [Ηλεκτρονικό]. Available: https://www.medcalc.org/manual/logistic_regression.php.

[6] A. Chakure. [Ηλεκτρονικό]. Available: https://towardsdatascience.com/decision-tree-classification-de64fc4d5aac.

[7] I. Rish. [Ηλεκτρονικό]. Available: https://www.cc.gatech.edu/~isbell/reading/papers/Rish.pdf.

[8] C.-J. L. Yin-Wen Chang. [Ηλεκτρονικό]. Available: http://www.jmlr.org/proceedings/papers/v3_old/chang08a/chang08a.pdf.